Lecture Notes on Language and Literature (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2023.061103
ISSN 2523-5869 Vol. 6 Num. 11

Theme Classification of the Complete Song Ci from the Perspective of the Digital Humanities

Yuanyuan Fang

Xihua University, Chengdu, Sichuan, 610000, China

Keywords: Digital humanities; the Complete Song Ci; topic classification; unsupervised learning

Abstract: To fully explore the underlying themes in the Complete Song Ci, we adopted a new paradigm of the digital humanities to efficiently extract themes from large-scale ancient poetry texts, which is expected to provide new perspectives and ideas for the study of traditional poetic themes. Under the BERTopic classification framework, we carried out fine-tuning training by combining a pre-training model for BERT with ancient Chinese and the SimCSE unsupervised learning method. We derived topic classification results of the Complete Song Ci through quantitative and visual means. The results indicate that the Complete Song Ci is divided into 43 sub-themes, among which certain similarities and compatibilities exist. After a further merging of the sub-themes based on cosine similarity values, we identified ten distinct themes, conforming to the Ten Major Themes theory of classical Chinese literature proposed in previous research, simultaneously establishing the research value of machine learning theories such as BERTopic in the topic classification of ancient poetic texts.

1. Introduction

With emerging digital trends, the production, availability, validity, and management of cultural materials have become issues posing new challenges for humanists[1]. Within this context, the interdisciplinary field of digital humanities has come to the fore. The rise of the digital humanities has contributed to the modern transformation of the humanities. Since the start of the 21st century, ancient Chinese literature has begun to reveal expansive focuses, among which classification and thematic studies on Tang and Song lyrics from a macroscopic standpoint are especially popular[2]. To classify classical poetry for thematic research, humanities scholars have traditionally adopted keyword searches and manual research methods. These efforts are time-consuming, labor-intensive, and suffer from the drawbacks of omission and fragmentation, not to mention a macro exploration of the landscape of Chinese literature through intelligent analysis. In recent years, the increasing maturity of topic modeling technology has provided a feasible path for automated topic extraction, in-depth mining, and analysis of poetic texts, thus helping humanities scholars to escape the needle-in-a-haystack process of literature searches, data mining, and text comparison. This is significant in enhancing the quality, depth, and impact of literary research in many ways.

Classical Chinese literature is as vast as the sea, and to facilitate the search for and citation of ancient materials, the idea of classification emerged as well as diverse perspectives on classification.

Themes are often referred to as the main content of literary works or the primary subject matter of literary creation. Classifying literary works according to themes is based on an in-depth understanding of the text. However, the following difficulties often arise in classifying the themes of literary works: EMSS.

• Crossover of themes: Influenced by context diversity and the ever-changing thoughts of the narrator, the fundamental themes of literature works tend to overlap. These themes are often intertwined with the expression of emotion in the writing of a scene, the expression of aspiration in the chanting of objects, and the chanting of emotions in chronicled events. This variety of themes has contributed to the enduring charm of literary works but has also left future generations with the difficult task of distinguishing themes.

• Mastering the main and secondary themes: The cross-fertilization of themes makes it difficult to distinguish between the main and secondary themes. For example, in Yanshu's Huanxi Sha – A New Song and a Glass of Wine, the phrasing is not only about the stunning scenery of the twilight spring, but also expresses the lament of the lyricist, who mourns the spring and cherishes the time and irrevocable beauty of things, so it becomes crucial to determine whether the overriding theme of the phrasing is 'spring scenery' or 'cherishing the time'. The subtleties of emotional expression often require scholars to spend a great deal of effort carefully distinguishing between primary and secondary themes.

• Uniformity of criteria for classifying themes: There are numerous themes to be explored; these include the weather, the seasons, animals, implements, emotions, religion, events, and festivals. This breakdown contains dozens of themes.

Since ancient times, there has been no fixed standard for the division of themes in classical literature; the reason for this is that the unity of the criteria for the division of categories has not been universally accepted. If a rough division were made, it would be easy to adopt a haphazard and disorganized approach; if an overly delicate distinction were made, it would be too elaborate and complicated. For these three reasons, there is still a long way to go in studying the themes of classical literature.

Thematic modeling techniques may help humanities scholars overcome the difficulties of thematic classification. This may offer broad prospects for reorienting ancient literature research to the era of big data. On the one hand, the rise of the digital humanities has gradually ushered text reading into an era of distance reading, where the automated processing and analysis of computers of massive amounts of data can facilitate the detection and observation of patterns that are too grand, sluggish, or complex for the human eye or brain to perceive and understand[3]. In adapting to this shift in perspective, humanities scholars have left behind overly limited in-depth analyses in favor of holistic, distant readings that are synergistic with the macro perspective. Although this inevitably sacrifices a certain degree of literary aesthetics, trends, relationships, and patterns based on the 'commonalities' of a large literature base, data from a large literary sample are more objective and scientific.

However, humanities scholars are not actively choosing and mastering appropriate computer technologies and computational thinking. They are also dovetailing and bidirectionally integrating traditional concepts of ancient literary studies, which are complementary rather than alternative[4]. Hence, using theme-modeling technology to achieve theme extraction is important for the innovative development of theme classification research in ancient literature. Song lyrics are among the brightest pearls of ancient literature. The Complete Song Ci, compiled by Mr. Tang Guizhang, is a masterpiece that collates song lyrics; it is also an indispensable source for studying them. Thus, we used a thematic model to classify the themes of the Complete Song Ci in the context of the digital humanities to provide a new perspective for studying song lyrics.

2. Related research

2.1 Theme studies on the Complete Song Ci

Themes in ancient literature are often linked to subjects and images. Centuries of chanting by poets and depictions by writers (of cuckoos, cicadas, crickets, willows, pines, and cypresses in flora and fauna, as well as changes in geography and celestial and climatic phenomena) have been imbued with certain symbolic meanings, showing and acquiring a specific thematic meaning[5]. The first is the study of individual imagery or themes, such as birds[6], rain[7], night[8], and paper[9] in all song lyrics, which are mostly combined with specific lyrical works to analyze the different emotional connotations and stylistic features of the target imagery in song lyrics. In contrast, studies of single categories of themes—including terms related to insects[10], birthdays[11], and festivals —usually reveal the deeper meanings of a song's social life and cultural patterns. Although they are slightly better in terms of the thickness of their research[12], they are mostly confined to discussions of specific phenomena and are still insufficient for exploring the themes of song lyrics. In addition, since the modern era, some scholars have focused on the problem of classifying song lyrics as a whole. Through collation, they found that when the lyrics were represented by Mr. Xu Boqing— who identified and categorized the Complete Song Lyrics Ci piece by piece and carried out multifaceted[13], multilevel literary interpretations—scholars eventually refined the classification into thirty-six categories. Based on all classical literature, Mr. Wang Li collected and analyzed works in various genres, including the Complete Song Lyrics Ci, and concluded that there are ten ephemeral themes in ancient Chinese literature: cherishing time; longing for each other; going away; wistfulness; grief for autumn; hatred of spring; wandering immortality; homesickness; mi-li; and life and death[14]. Owing to the limitations of technological developments in their time, none of these scholars introduced computer technology to help summarize the themes of the Complete Song Lyrics Ci, and they still adopted the traditional paradigm of literary research without combining it with additional visualization tools.

By conducting literature and online research, we found that manually reviewing and collating the themes of the Complete Song Lyrics Ci have become increasingly prevalent. However, given that the use of theme modeling technology to extract the themes of the Complete Song Lyrics Ci is a pioneering approach in the field of ancient literature research, and that no research has yet used theme models under deep learning paradigms such as BERTopic, we selected BERTopic theme modeling technology to carry out a study on the themes of the Complete Song Lyrics Ci. We adjusted the perspective of traditional theme research on song lyrics, which is no longer limited to the in-depth analysis of a single theme, but uses the theme modeling technique for a broader exploration of motifs in ancient literature. This will help provide more regular, macroscopic, and trendy possibilities and clues for studying song lyrics as a whole.

2.2 Thematic model studies

Topic models are a popular and effective technique in text mining and are often used to explore potential topics in a text. Latent Dirichlet allocation (LDA) is a classical topic model. The LDA topic model proposed by Blei, Ng and Jordan is widely used for text classification, text modeling, image processing, and information retrieval[15-16]. For example, Zizhuo, Ying and Yanqiu used the LDA topic model to mine poems about classical musical instruments in All Tang Poems and All Song Lyrics[17], providing a brand new analytical approach to the study of musical instruments in classical literature. Kaiyan, Yao and Qianused an LDA topic model to explore the subordination or relevance of botanical imagery to themes in contemporary Chinese literature based on themes and terms that had already been identified. Based on the text resource People's Daily[18], Qi used a combination of the LDA topic model to extract keywords and high-frequency terms to interpret changes in social life across different

periods through newspaper vocabulary[19] , which is conducive for distant reading. For example, in the process of generating topics, the traditional LDA topic model and its derivative models employ the bag-of-words idea to extract topic vocabulary, which cannot effectively represent the semantic and grammatical order relationships between words, and ignores the role of low-frequency words in the text corpus, making the modeling results less interpretable and difficult to determine. It is challenging to decide on an optimal number of topics[20]; therefore, it is difficult to accurately understand topics in ancient literary works. To solve the problem of the inaccuracy of bag-of-words-based topic models in text modeling, Mikolov et al. proposed Word2Vec in 2013[21], a topic classification model rooted in the idea of word embedding, which involves embedding a high-dimensional lexical space into a low-dimensional continuous vector space. Another example is Top2vec[22], which is a topic model grounded in word embedding that measures the similarity of topics by calculating the vector similarity, unlike the bag-of-words method of randomly generating topic words. This type of topic model produces subjects that are primarily based on density clustering; it captures topic words mainly by judging whether they are located in the center of the mass of the topic aggregation region. In short, the closer the word-embedding vector is to the center of mass of the aggregation, the more the corresponding vocabulary is prioritized as the theme word of the region. Although the embedded topic model is more efficient and general and can avoid the disadvantages of randomness and the neglect of deep semantic relationships, it is still not entirely suitable for solving cross-cutting challenges and primary subordination in humanities research. This represents multiple words in different contexts as the same vector, which cannot solve the problem of semantic clustering, thus making the results of literary research prone to bias. When faced with a large volume of text data, the results based on density clustering are too dense, which can easily lead to a misjudgment of the subject terms.

To solve the aforementioned problems, Devlin et al.[23]proposed the BERT model in 2018 to bridge the gap between related models such as Word2vec. BERT is a deep bidirectional representation pre-training model rooted in contextual information that combines the training of a large-scale corpus to generate word vectors that fully integrate context semantics, allowing for the dynamic implementation of multi-sense words in different contexts. BERTopic is a thematic model derived from BERT[24], which has opened new horizons for natural language processing (NLP) techniques. This method can be applied to most language models. The algorithmic tools are selected and combined with the user's needs and actual resources to fulfill the user's requirements and fine-tune the pre-trained model to transform the text into a more accurate document vector. As such, BERTopic is less likely to ignore the order between words in the generation of topics. This process is grounded in density-based clustering. However, to select topic words, BERTopic chose the C-TF-IDF algorithm to extract keywords for each topic category, thereby bridging the gap between probability-based topic generation and central sampling under density-based clustering.

Furthermore, with the recent strong introduction of the multi-modal macro model GPT-4, ChatGPT (based on the GPT-4 model)—which has driven a round of revolution in the field of NLP—gave the following answer after it was asked if it could quickly classify the topics of the Complete Song Lyrics Ci[25]:

The Complete Song Lyrics Ci is a collection of words written by lyricists during the Song Dynasty. Based on their content, we can classify these works according to themes such as landscapes and gardens, frontier garrisons, parting and longing, festivals and customs, nostalgia and sadness, history, banquets and leisure, and friendship and travel. This is only a partial classification of themes found in the Complete Song Lyrics Ci. It is important to note that many lyrics deal with more than one theme, and there may be crossover and fusion between these themes.

As can be seen, ChatGPT can initially roughly classify the themes of a work's text. However, it is not comparable to a dedicated theme model in terms of the granularity of classification, and the

training cost of ChatGPT is extremely high compared to that of a dedicated theme model. Although representative words can be listed under each classification, it is feasible to obtain a comprehensive understanding of all classified words. ChatGPT also highlights the possibility of crossover and convergence among topics, underscoring the difficulty of topic classification and the need for topic models.

3. Technical routes

After collating the relevant studies mentioned above, we launched an in-depth analysis of all words from the Song Dynasty based on the BERTopic theme classification framework. We aimed to examine how to quickly classify topics among a certain class of Song words in humanities computing, to achieve large-scale text extraction and classification, and to quickly find similar high-frequency words for that class of topics to provide more interpretability of the text in quantitative processing. The main methodological routes are as follows (see Fig. 1).

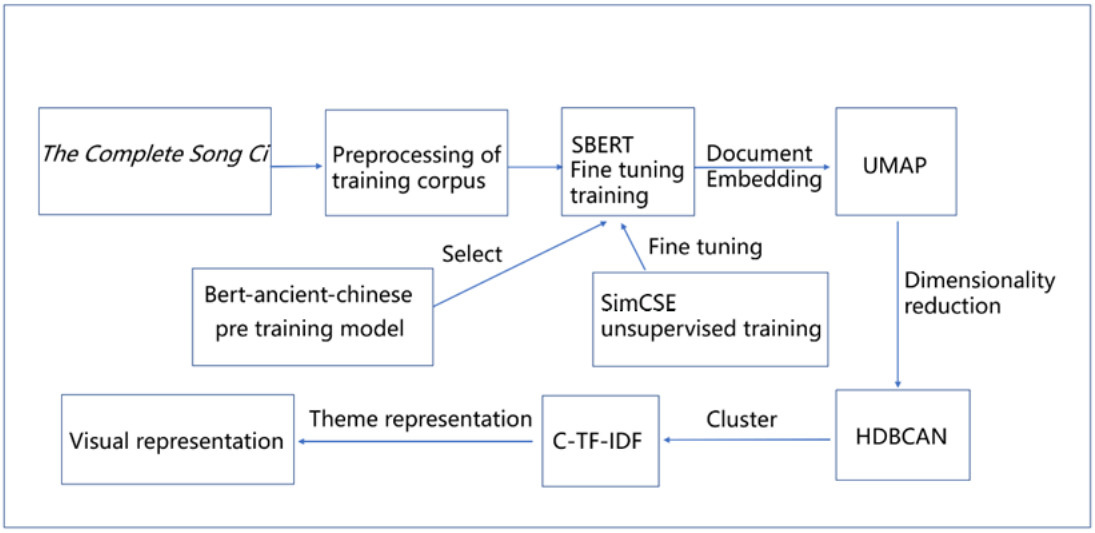

Figure 1: Schematic diagram of the BERTopic topic extraction process

3.1. Data selection and preprocessing

First, the Complete Song Lyrics Ci were compiled by Mr. Tang Guizhang to form a database, and data cleaning and other pre-processing work were carried out on this text to further improve the quality of the data, guaranteeing the effectiveness of the data calculation. In particular, all names of authors, word names, missing words, and unnamed words were removed from the Complete Song Lyrics Ci, and only the text of the Song lyrics, which express complete meaning, was retained. After cleaning the data, 18,979 valid lyrics were classified. Second, the cleaned Song lyrics were divided by applying jiayan, a partitioning tool in the field of ancient Chinese books, and removing the stop-words to improve the accuracy of the subsequent model classification.

3.2. Pre-training and fine-tuning

Converting text words into spatial vectors is one of the core steps of the BERTopic topic classification framework, which uses sentence-BERT (SBERT) to train and learn all Song words to achieve the conversion of computer-recognizable vectors[26] . SBERT is a good representation-based text-matching model with good practicality at present, and by modifying the original BERT network with a twin network structure, it can output fixed-length sentence vectors that retain semantic

information. Although a Chinese version of the official SBERT pre-training model was provided, the results obtained directly using this Chinese model to extract the themes of all Song words were not satisfactory. Thus, a pre-training model for the automatic processing of ancient texts, BERT-ancient-English[27], was introduced and fine-tuned for the all-Song theme classification task. However, due to the large scale of the text data volume and the overwhelming task of manual data annotation, to further improve computational efficiency and explore unsupervised fine-tuning learning for other similar studies simultaneously, we used SimCSE for unsupervised training to obtain richer text learning features[28]. The two steps above—that is, the selection of the ancient text pre-training model and unsupervised fine-tuning training—significantly improved the performance of the topic model, which was more adaptable to the topic classification task of all Song lyrics.

3.3. Dimensionality reduction and clustering

Dimensionality reduction and clustering of the vector results are the most important steps in obtaining thematic classification outcomes. UMAP is a recently proposed dimensionality reduction technique based on stream learning[29]; it can preserve the global structure of the data and achieve effective dimensionality reduction by constructing the key structure of vectors in a high-dimensional space and mapping them to a low-dimensional space. We employed UMAP to reduce the dimensionality of the all-Song vector to remove redundant features and quickly identify the core features of the text. The clustering process can be simplified into the following steps[30]: estimating density, selecting high-density regions, merging points in these selected regions, and extracting them into clusters. After clustering, we obtained multiple clusters, each representing a theme in all Song lyrics.

3.4. Visualization and literary interpretation

After automatically extracting the topic words of each cluster using the C-TF-IDF keyword extraction algorithm, we visualized the results of the thematic classification of the Complete Song Lyrics Ci. Aggregating and refining the data visualization results can help identify extremes, anomalies, trends, and regularities in the thematic classification of the Complete Song Lyrics Ci, thus providing a comprehensive understanding of the literature.

4. Analysis of the results of the thematic classification of all Song words

After conducting experiments according to the above technical lines, we completed the extraction and classification of the themes of all Song words, and we the obtained results as follows:

After subject extraction of more than 18,000 Song Dynasty lyrics, we identified 43 sub-themes. Each sub-theme contains a certain number of classified song lyrics. After clustering the themes of all Song lyrics, we derived the theme clustering distribution map (see Fig. 2), in which dense areas are distributed in various colors, meaning that Song lyrics of the same theme are located near each other. The size of the sub-themes can be observed intuitively according to the size of the color range. Based on the principle of density clustering, if an area is not sufficiently dense, the topic model does not determine the text in that area as a cluster of topics that can be clustered. Take Top2vec, a topic model based on word embedding, as an example: When using less complex English-language news text for topic clustering, there is also a situation in which the text topic cannot be determined. Thus, it is reasonable and customary for the BERTopic topic model to include unidentifiable texts when pinpointing issues. Furthermore, the bottom 0 theme had the most prominent color. The Song lyrics in this theme are mainly about landscapes, so the articles about landscapes in all Song lyrics comprise a more significant proportion.

Figure 2: Theme clustering distribution of all Song lyrics

After merging several sub-themes based on hierarchical clustering, we derived ten categories of major themes in all Song words. Per topic does not directly formulate the name of the article after extracting the music; we composed the words already categorized under each theme and grasped its main characteristics by combining the core theme words given under each sub-topic to formulate the name of the theme. However, due to the excessive number of sub-topics generated and because there are many similar sub-topics, to further accurately grasp the content of the topic, and to simplify the content of similar sub-topics, we can sub-topics; hierarchical clustering can solve this problem and the resulting application method. Hierarchical clustering is based on calculating the similarity of document vectors for each sub-topic, which helps to find two more similar clustered topics and merge them hierarchically. Thus, hierarchical clustering dendrograms reflect the aggregation of the same topics and make topic extraction more cohesive. After the hierarchical clustering of all the sub-themes of all Song words, we can see that BERTopic categorizes the themes of all Song words into ten major categories (see Fig. 3). These ten categories of articles are organized, and the proposed theme names are as follows: spring, festival, love, boudoir, autumn, Buddhist, birthday, wistful, wandering fairy, and idle.

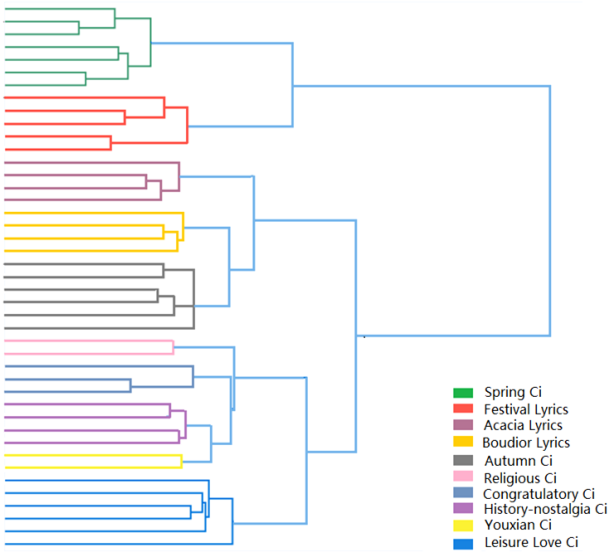

Figure 3: Cluster diagram of the thematic level of all Song words

BERTopic's extraction and classification of themes in the Complete Song Lyrics Ci are in line with literary logic and can be promoted in classical literature research. BERTopic's theme classification of the Complete Song Lyrics Ci has a specific theoretical foundation. In the late 1980s, under the influence of Western literary theories, the study of Chinese classical literature further developed under the intersection and collision of Chinese and foreign trends, among which literature eventually evolved into a trend of epi-phenomenology, with scholars represented by Wang Li and Tan Guilin, who further promoted the formation of the theoretical system, methods, and models of thematic studies. Wang Li applied mathematics to study ancient literature, proposing that classical Chinese literature contains ten major themes and nine central images, revealing the mechanism of literary production and the operation more three-dimensionally. This viewpoint summarizes classical literature's different aesthetic connotations and functional effects, reflecting the near and far views of the spirit, mind, and culture. In digital humanities research, the 'black box' of algorithms is key to vigilance among humanities scholars. When using computer tools, humanities scholars must analyze the computing principles and judge whether the results are objective. The 'ten themes' of early Chinese literati pointed out by Mr. Wang Li are based on the broad scope of Chinese classical literature, while this paper captures the main features of the text based on the BERTopic theme classification framework, and extracts and merges the themes of all Song lyrics. The ten pieces of Song lyrics are shown in Fig. 3, among which some articles are classified as 'spring lyrics,' 'autumn lyrics,' 'acacia,' 'youxian,' and 'huaigu.' These articles coincide with Mr. Wang Li's ten themes, but they are also based on lexical clustering. The themes of the text are more precisely extracted and refined, presenting the unique aspects of Song lyrics, which are derived from Chinese classical literature, such as 'leisurely feelings.' The category of 'leisurely feelings' shows that Song lyrics are different from Tang poems and Song poems in that they are more concerned with the state and reality, and their leisurely content and lyrical qualities express the leisurely life of Northern Song lyricists. This also fully reflects the actual feasibility of the BERTopic theme model for classifying literary works.

BERTopic can present the core keywords for different categories of themes in all Song Lyrics. Here, we selected the core words of sub-themes 0–7 due to space limitations (see Fig. 4). Through comparison, we found that (except for sub-theme 4) based on the keywords of 'spring,' 'flower,' 'mountain,' and 'water,' the other sub-themes contained very similar content. We selected the core words of sub-themes 0–7 (see Fig. 4) through comparison, except for sub-theme 4, which is derived from the keywords of 'spring,' 'flowers,' 'swallows,' 'mountains,' and 'water.' The other sub-themes cover very similar content, mostly natural spring scenes. Therefore, we classified these sub-themes as 'spring lyrics.' In categorizing the themes of 'all Song lyrics,' we replaced the previously adopted pieces of 'hate of spring' and 'autumn lyrics' with 'spring lyrics' and 'autumn lyrics,' which refer to a broader range. The reason for this is that the theme classification of 'spring lyrics' and 'autumn lyrics' is based on the 'spring lyrics' and 'autumn lyrics.' The reason is that the sub-themes of 'spring lyrics,' for example, cover seasonal features related to 'spring' such as 'cold'; natural phenomena related to 'spring' such as "flowers,'; social life related to spring, such as 'articles,'; and emotional expressions arising from spring According to the core keywords displayed in Figure 4, the word 'hate' appears only in sub-theme 7. The other sub-themes of spring words are often full of words that express the leisure and sorrow of spring, pleasure, and wilderness, but rarely hate and sorrow. The theme classification of 'spring hatred' and 'sadness for autumn,' which was adopted by previous writers, has been replaced by 'autumn words,' which is more reasonable. This shows that the attributes of spring and autumn are complex and subtle in song lyrics, that the richness of the spring and autumn scenery as well as the suitable climate can fully stimulate the passion and talent of the lyricists. Moreover, attention to spring and autumn reflects the lyricists' strong sense of time and life. Some value judgments on themes under contemporary consciousness can also partially reveal the underlying

ideological themes and their deeper meanings that the author and his predecessors have failed to realize. Hence, the cultural significance of ancient literature is evident. For example, modern society has advanced technological tools to help predict the weather. However, people are less able to perceive changes in the natural climate and their states of mind. In ancient times, however, when productivity was low, the ancients were especially sensitive to seasonal changes, especially the warm spring and the cold autumn, which provided a more suitable environment for the literati to create their works, grasp the passage of time in the gradual changing of the seasons, and understand subtleties, thus rendering emotions under complex and changing seasonal themes.

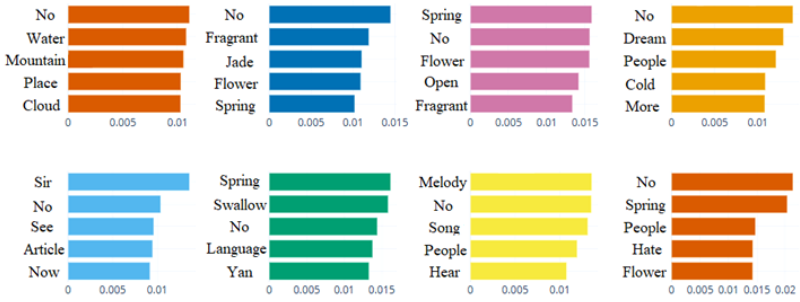

Figure 4: The core characteristic words of some themes of all Song words

Some themes in all Song lyrics were incompatible, whereas most were often incongruent. As mentioned earlier, owing to the cross-cutting themes among the pieces, it was difficult to delineate them. To better compare the differences among the themes of all Song words, we drew a heat map of theme similarity by calculating the similarity of each sub-topic document vector (see Fig. 5), where the shade of color blocks indicates the similarity values among sub-themes, and the similarity increases gradually from light to dark. The similarity heat map shows the similarities between sub-themes intuitively and efficiently. This helps researchers quickly find the connections between sub-themes and locate sub-themes with apparent differentiation. Looking at Fig. 5 carefully, we can see that there are different degrees of compatibility among sub-themes, and a sub-topic has both closely related sub-themes that have a higher degree of compatibility and more distant sub-themes that have a lower degree of compatibility. For example, sub-theme 28 is closely related to ‘boudoir’ and ‘separation.’ It tends to be ‘literature of the heart,’ focusing on the separation between individuals, so it has a certain degree of compatibility with other sub-themes. The color block is darker, except for the wistful words belonging to sub-theme 34, which are less compatible, and the color block is lighter. Second, sub-themes 13, 14, 22, 35, and 41 were exceptional. These sub-themes split the similarity heat map and differ significantly from other sub-themes, among which sub-theme 35 encompasses the extreme values of the entire similarity heat map and forms a typical representation of research differentiation. These distinctly differentiated sub-themes are ‘special themes’ that are not easily integrated compared to other sub-themes, and their differentiation highlights the fact that the sub-themes they represent show more significant and precise, unique, and easily identifiable characteristics. For example, sub-themes 13 and 14 are festival words that do not include all festivals, but include two festivals: the mid-autumn festival, and the Chung Yeung Festival. The reason why these two sub-themes do not easily blend is that the core keywords, such as ‘mid-autumn’ and ‘Double Ninth Festival’, highlight the season in which the words were written and are not necessarily associated with spring words, which account for a large proportion of the words in all Songs, so the color block of the heat maps of these two sub-themes is lighter when it is associated with spring lyrics. However, because the season of composition is in autumn, the color of the block is slightly darker

when compared with the sub-themes representing autumn lyrics, and there is partial compatibility. Finally, as in the case of sub-theme 35, it had the weakest similarity in the heat map overall. According to the hierarchical clustering chart, sub-theme 35 belongs to the category of wandering immortal words, which mostly depict the 'immortal realm' as the central content theme of Song lyrics, or use wandering immortals to express their longing for immortality and thus their resentment against reality. Due to its unique creative content, it has poor compatibility with other themes. In short, the compatibility between themes is either hindered by objective factors, such as the conflict of seasons, or needs to be narrower in content, and the subject matter must be compatible.

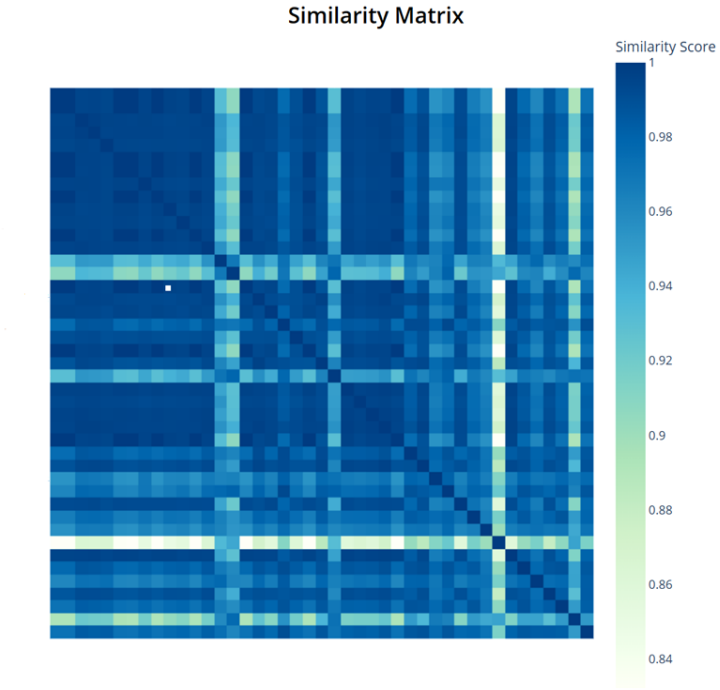

Figure 5: Heat map of the similarity of themes in all Song words

The genre boundaries of song lyrics are apparent, as the styles of euphemism and bold schools are different. Hence, the objects chosen for depiction are other ones, and the themes of the song lyrics are also different. Based on the core keywords of the articles presented in the above processing results, many genre-specific words can be seen directly in the figure. The sub-themes related to the objects depicted by the graceful and restrained poetic genre have a high degree of compatibility, such as 'lovesickness' and 'dreaming back,' which precisely reflects the artistic style of the euphemists, who were concerned with the pain of parting of women and wanderers, and showed the artistic style of the euphemists. On the contrary, the euphemists were concerned with 'heroes,' 'laughter,' and 'hunting.' In contrast, the bold and unrestrained lyricists focused on 'heroes,' 'laughter,' and 'hunting,' and the degree of compatibility of these sub-themes is also higher. Song lyrics can be more clearly and intuitively identified from the standpoint of the digital humanities, which is precisely the innovative perspective that the digital humanities bring to the field of Song lyric analysis. By comparing the similarities between themes, it is possible to find pieces that are easily integrated into ancient literature and less popular articles. Through the connections between themes, it is also possible to explore alternative themes in old literary works, to think outside of the box, and to holistically interpret classical literary pieces from multiple angles.

5. Conclusion

Employing the framework of the BERTopic theme model, we classified and extracted themes from the masterpiece of the Song lyrics collation, all Song lyrics. We analyzed the distribution pattern of articles, the characteristic core words within the pieces, and the similarity status among themes in all Song lyrics. From the perspective of the digital humanities, reading from afar and examining from near, the theme model under the deep learning paradigm represented by BERTopic is introduced into the theme study of ancient Chinese poetry. This is not only innovative in terms of the method, but also feasible in terms of literary analysis. This study is a valuable exploration and attempts worthy of being extended to more theme classification problems of ancient books. At the same time, during the study process, we found that the most significant challenge of the BERTopic-based theme study of all Song lyrics was the ambiguity of some Song themes. The classification and identification of literary themes are still difficult for manual recognition, let alone automatic classification via unsupervised computer learning. Currently, there is much room for improvement in the development of computer technology. In the future, in NLP, more attention should be paid to the accuracy of deep semantic recognition technology. Through the continuous optimization and improvement of algorithms, computers can further improve their ability to learn and understand human languages and words. This can also promote theme research in the traditional humanities, from the local to the macro levels, from subjective judgment to objective analysis, and from qualitative research to quantitative analysis, to further promote the deep excavation and innovative development of digital humanities research.

References

[1] Burdick A., Drucker J. and Lunnefeld P. Digital Humanities: Changing the Game of Knowledge Innovation and Sharing [M], Ma, Linqing Han, Ruohua, Translation. People's University of China Press, Beijing, 2018.

[2] Zhaopeng W. Progress and prospect of lexicographic research since the new century [J]. Academic Research, 6: 143–151, 2015.

[3] Börner K. Plug-and-play microscopes [J]. Communications of the ACM, 2011, 54(3).

[4] Zhaopeng W. and Dawei S. The initial practice and academic significance of digital humanities in studying ancient literature [J]. Chinese Social Sciences, 8: 108–129, 206–207, 2020.

[5] Daiyun Le(Ed). A Course on Comparative Literature between Chinese and Western [M]. Higher Education Press, Beijing, 1988.

[6] Xuehui Z. The art of bird imagery in song lyrics [J]. Journal of Soochow University (Philosophy and Social Science Edition), 32(2): 158–163, 2011.

[7] Ning X. Research on Rain Imagery in Song Lyrics [D]. Dissertation, Nanjing University of Information Engineering, 2022.

[8] Yuxuan Z. Study on the Imagery of Paper in Song Lyrics [D]. Dissertation, East China Normal University, 2021.

[9] Yixuan D. Study of Night Imagery in Song Lyrics [D]. Dissertation, Hunan University, 2016.

[10] Qian W. The development and metamorphosis of insect words in the Song Dynasty from butterfly and cricket words [J]. Jiangsu Social Science, 4: 228–232, 2014.

[11] Yang L. Aria of life and talent: Aesthetic description of the creation of Song Dynasty shou lyrics [J]. Masterpiece Appreciation, 6: 34–37, 1995.

[12] Gai H. Research on Song Dynasty festival words: A literature review [J]. Chongqing Social Science, 2: 78–83, 2013.

[13] Boqing X. Studies on Song Lyric Themes [M]. China Book Bureau, Beijing, 2007.

[14] Wang L. Ten Themes of Ancient Chinese Literature: Archetypes and Fluxes [M]. Liaoning Education Publishing House, Shenyang, 1990.

[15] Blei D. and Jordan M. I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.

[16] Li X., Hu Y. and Huang L. A study on hybrid automatic classification of multiple types of documents using LDA topic model. Library Forum, 35(1): 74–80, 2015.

[17] Zizhuo S., Ying Y. and Yanqiu S. Topic model-based text mining of classical musical instrument poems [J]. Journal of Chinese Information, 33(3): 79–86, 2019.

[18] Kaiyan M., Yao X. and Qian C. A study of plant imagery in contemporary Chinese literature in the digital humanities perspective. Digital Humanities Research, 2(2): 35–45, 2022.

[19] Qi L. The spirit of the times: Keyword extraction and interpretation of short texts – Practice based on the text of People's Daily [J]. Digital Humanities, 3: 125–150, 2020.
[20] Zhang D. X. and Zhang M. A review of the progress of research on applying the LDA topic model in the field of graphical intelligence [J]. Library Intelligence Knowledge, 39(6): 143–157, 2022.
[21] Mikolov T., Chen K., Corrado G. and Dean J. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv: 1301. 3781, 2013.
[22] Angelov D. Top2vec: Distributed representations of topics [J]. arXiv preprint arXiv: 2008. 0947, 2020.
[23] Devlin J., Chang M. W., Lee K and Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 1810. 04805, 2018.
[24] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv: 2203. 05794, 2022.
[25] OpenAl. GPT-4 technical report [J]. arXiv preprint arXiv:2303. 08774, 2023.
[26] Reimers N. and Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT networks [J]. arXiv preprint arXiv:1908. 10084, 2019.
[27] Wang P. and Ren Z. The uncertainty-based retrieval framework for ancient Chinese CWS and POS [J]. Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages. 164–168, 2022.
[28] Gao T., Yao X. and Chen D. SimCSE: Simple contrastive learning of sentence embeddings [J]. arXiv preprint arXiv:2104. 08821, 2021.
[29] McInnes L., Healy J. and Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802. 03426, 2018.
[30] McInnes L. and Healy J. Astels S. hdbscan Hierarchical density-based clustering [J]. Journal of Open Source Software, 2(11): 205, 2017.